3M

УДК 681.3

А.А. Марченко
Киевский национальный университет имени Тараса Шевченко, Украина
Украина, 03680, г. Киев, пр. Глушкова 4-д,

Алгоритм автоматического определения семантических отношений между концептами онтологий

O.O. Marchenko
Taras Shevchenko National University of Kyiv, Ukraine
Ukraine, 03680, c. Kyiv, Glushkova Ave., 4-d

Algorithm for Automatic Determination of Semantic Relations Between Concepts of Ontologies

О.О. Марченко
Київський національний університет імені Тараса Шевченка, Україна
Україна, 03680, м. Київ, пр. Глушкова, 4-д

Алгоритм автоматичного визначення семантичних відношень між концептами онтологій

В статье описан метод автоматического определения семантических отношений между концептами-узлами сети онтологической базы знаний на основе анализа матриц семантико-синтаксических валентностей слов. Данные матрицы получены при помощи неотрицательной факторизации тензоров синтаксической сочетаемости слов. Тензоры были сгенерированы в процессе частотного анализа синтаксических структур предложений текстов статей English Wikipedia.
Ключевые слова: обработка текстов на естественном языке, неотрицательная тензорная факторизация, онтологии.

This paper describes a method for automatic determining semantic relations between concept nodes of an ontological knowledge base by analyzing the matrices of words semantic-syntactic valences. These matrices are obtained by non-negative factorization of tensors of words syntactic combinability. The tensors have been generated in the course of syntactic structures frequency analysis for the large text corpus of English Wikipedia articles.
Key words: natural language text processing, non-negative tensor factorization, ontologies.

У статті представлений метод автоматичного визначення семантичних відношень між концептами-вузлами мережі онтологічної бази знань на основі аналізу матриць семантико-синтаксичних валентностей слів. Дані матриці отримані невід'ємною факторизацією тензорів синтаксичної сполучуваності слів. Тензори були сформовані в процесі частотного аналізу синтаксичних структур речень текстів статей English Wikipedia.
Ключові слова: обробка текстів на природній мові, невід'ємна тензорна факторизація, онтології.

Вступ

Неотрицательная тензорная факторизация в последнее время широко востребована в таких областях, как информационный поиск, обработка изображений, обработка естественного языка, машинное обучение и в других смежных направлениях. Данный

подход является одним из наиболее перспективных для выявления и анализа взаимосвязей, и отношений в данных, где сочетаются объекты N разных типов и классов. N-мерный тензор, который в информатике трактуется как *многомерный массив данных*, является удобной структурой для представления данных высших порядков. Факторизация N-мерного тензора генерирует N матриц, состоящих из k векторов, которые представляют отображение каждого измерения тензора на k факторизованных измерений скрытого семантического пространства, что служит уникальным средством для моделирования и выявления взаимосвязей и совместного поведения N переменных в массиве N-мерных данных. Факторизация тензора является мультилинейным аналогом сингулярного разложения матриц, используемого в латентном семантическом анализе для обработки двумерных массивов данных. В некотором смысле метод неотрицательной тензорной факторизации тензоров можно назвать n-мерным обобщением латентного семантического анализа.

В настоящее время неотрицательная тензорная факторизация является перспективным методом в решении задач компьютерной лингвистики, о чем свидетельствуют многочисленные работы в этом направлении [1-4].

Данная робота описывает модель многомерного представления семантико-синтаксических отношений между словами в предложениях естественного языка. Частотный анализ структур предложений большого текстового корпуса дает описание естественного языка в виде многомерного массива возможных сочетаний слов в определенных синтаксических позициях, которые и задают данный язык. Многомерный разряженный тензор раскладывается с помощью метода неотрицательной факторизации, который, помимо компактной и удобной структуры представления данных о сочетаемости последовательностей лексем в некоторых синтаксических позициях предложений естественного языка, дает эффективный метод вычисления оценки вероятности существования семантико-синтаксических связей между словами разных грамматических категорий. Каждому слову в разложенных матрицах тензора соответствуют вектора уменьшенной размерности k (где k – размерность латентного семантического пространства разложенного многомерного лингвистического тензора), и эти вектора описывают семантико-синтаксическое поведение данного слова: в какого типа связи и с какими словами оно вступает. По аналогии с химией можно рассматривать k-мерные вектора слов из матриц разложенного тензора как векторы семантико-синтаксических валентностей – (semantic valence vector – **SVV)** слов. Слова по своей природе являются неоднозначными, и одному слову, как правило, соответствует несколько значений. В работе предложено рассматривать k-мерные **SVVs** слов как суммы составных слагаемых **SVVs** разных значений этих слов. В статье представлен разработанный метод расщепления **SVVs** слов на составные слагаемые **SVVs** их разных значений и способ привязки этих расщепленных составных слагаемых **SVVs** к синсетам WordNet [5] в качестве их собственных значений **SVVs**, неявно описывающих их семантические отношения с другими синсетами WordNet.

Модель N-мерного пространства представления семантико-синтаксических отношений слов в предложениях ЕЯ – текстов

Рассмотрим пример некоторого текстового корпуса «Футболист забил гол. Девочка ела мороженое. Футболист забил гол.», состоящего из трех предложений. Первое и

третье предложение совпадают. В трехмерный массив записываются частотные оценки сочетаний слов в позициях ***подлежащее***, ***сказуемое*** и ***дополнение***.

V(Футболист, забил, гол) = 2
V(Девочка, ела, мороженное) =1
Представим трехмерный массив в виде трехмерного пространства, где оси OX, OY и OZ соответствуют синтаксическим позициям слов – ***подлежащему***, ***сказуемому*** и ***дополнению:***
Ось OX(подлежащее) (Футболист, Девочка) (x1, x2)
Ось OY(сказуемое) (забил, ела) (y1,y2)
Ось OZ(дополнение) (гол, мороженное) (z1,z2)

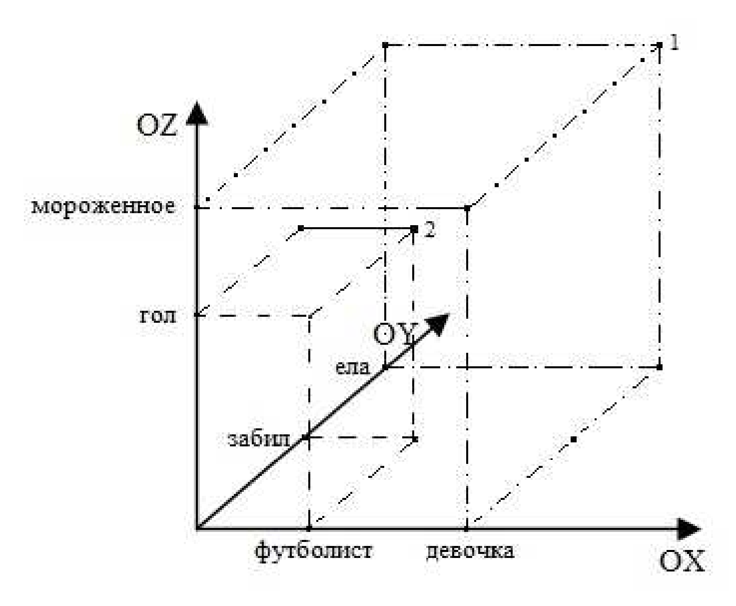


Рисунок 1 – Графическое представление трехмерного семантико-синтаксического пространства слов, моделируемого трехмерным тензором

Данный набор предложений можно попытаться промоделировать системой уравнений:

1. x1*y1*z1=2
2. x2*y2*z2=1
3. x1*y1*z2=0
4. x1*y2*z2=0
5. x1*y2*z1=0
6. x2*y1*z1=0
7. x2*y1*z2=0
8. x2*y2*z1=0

Очевидно, что система решения не имеет.
Поэтому x1, y1, z1, x2, y2, z2 целесообразно представить в виде сумм k переменных:

$$\sum_{t=1}^{k} x_{it} * y_{jt} * z_{lt} = 2, \text{ если } i=j=l=1;$$

$$\sum_{t=1}^{k} x_{it} * y_{jt} * z_{lt} = 1, \text{ если } i=j=l=2;$$

$$\sum_{t=1}^{k} x_{it} * y_{jt} * z_{lt} = 0, \text{ если не } i=j=l.$$

Одно из возможных решений системы уравнений, где k=2,
описано в виде матриц X, Y и Z:

$$X$$
$$\begin{pmatrix} & x_1 & x_2 \\ level_1 & 2 & 0 \\ level_2 & 0 & 1 \end{pmatrix}$$

$$\mathrm{Y}$$
$$\begin{pmatrix} & y_1 & y_2 \\ level_1 & 1 & 0 \\ level_2 & 0 & 1 \end{pmatrix}$$

$$\mathrm{Z}$$
$$\begin{pmatrix} & z_1 & z_2 \\ level_1 & 1 & 0 \\ level_2 & 0 & 1 \end{pmatrix}.$$

В матрицах X, Y и Z уровни, соответствующие рядкам, можно представить семантико-тематическими подпространствами: первый уровень соответствует футбольной тематике, второй – теме «питание, продукты питания».

При обработке крупных текстовых корпусов большого объема мы получим модель представления правильных 3-словных предложений естественного языка. При этом количество строк матриц будет значительно меньше, чем количество разных сочетаний слов (в приведенном выше примере 2 разных сочетания – 2 строки). При построении данной модели используется принцип, аналогичный сингулярному разложению матриц в латентном семантическом анализе [6].

В ЛСА разряженная матрица D большой размерности $N\times M$, хранящая оценки частоты использования слов-терминов в разных текстах корпуса (N – количество слов терминов, а M – количество текстов) раскладывается на две матрицы A $(N\times k)$ и B $(k\times M)$, где k намного меньше N и M. Матрица A представляет собой отображение множества терминов-слов в k-мерное пространство латентных фактор-признаков, а матрица B – отображение множества текстов в это пространство. Можно представить себе каждое измерение этого пространства в виде некоторой тематики, по которой коммутируются группы ее слов-терминов и тексты данной направленности. Схематически понятийную интерпретацию результатов работы метода можно представить, как показано на рис. 2.

Структуру, полученную в результате работы алгоритма латентного семантического анализа, можно сравнить с трехслойной нейронной сетью. Данная сеть состоит из двух слоев, которые представляют множества объектов двух типов, а также из скрытого коммутационного слоя, состоящего из множества узлов с различными весовыми коэффициентами. Этот слой моделирует взаимосвязи между этими объектами двух типов и связывает данные двух слоев в единую нейронную сеть.

Отдельного рассмотрения заслуживает проблема выбора k при разложении разряженной матрицы большой размерности D. Самым идеальным является случай, когда размерность модели латентного семантического пространства k (в нашем случае –

количество тематик) известно заранее. Иначе приходится определять это число автоматически при помощи специальных алгоритмов. При этом если происходит ошибка в сторону уменьшения числа *k* – некоторые разные тематические измерения латентного семантического пространства сольются в одно измерение. Если же число *k* будет больше реального количества тематик в текстах корпуса, то некоторые тематические измерения пространства будут искусственным образом разделены на несколько измерений. И в первом, и во втором случае качество модели будет падать тем больше, чем больше расчетное *k* будет отличаться от реальной тематической размерности текстового корпуса.

Помимо экономного представления разряженной матрицы *D*, получаем удобный инструмент для измерения семантической близости между словами-терминами, между текстами и также между словами и текстами. Для того чтобы определить значение меры близости между двумя словами, нужно вычислить скалярное произведение соответствующих им векторов строк матрицы *A*. Для того чтобы определить значение меры близости между двумя текстами, нужно вычислить скалярное произведение соответствующих этим текстам векторов столбцов матрицы *B*. Чтобы определить значение меры близости между словом и текстом, нужно вычислить скалярное произведение вектора строки этого слова из матрицы *A* и транспонированного вектора столбика этого текста из матрицы *B*.

Данный подход применен в представлении модели сочетаемости слов в предложениях естественного языка, описанной в начале работы. Для разложения трехмерного массива сочетаемости слов в предложениях структуры «подлежащее-сказуемое – дополнение» используется метод неотрицательной факторизации тензоров NTF [1].

Результат разложения разряженной матрицы D (термины×тексты) в виде произведения двух матриц A(N×k) и B(k×M) с уменьшенным k, выполненного методом латентного семантического анализа

Таблица 1

Матрица D	Собака Баскервилей	Пляшущие человечки	Властелин Колец	Хоббит	Искусство Программирования	Программа = Алгоритм + Структура данных
Расследование	X	X	0	0	0	0
Убийство	X	X	0	0	0	0
Похищение	X	X	0	0	0	0
Маг	0	0	X	X	0	0
Эльф	0	0	X	X	0	0
Гном	0	0	X	X	0	0
Орки	0	0	X	X	0	0
Оператор	0	0	0	0	X	X
Цикл	0	0	0	0	X	X
Процедура	0	0	0	0	X	X

Раскладывается в произведение матриц *A* и *B* (*X* – некоторые значения, отличные от нуля).

Таблица 2

Матрица А	Dim1	Dim2	Dim3
Расследование	X	0	0
Убийство	X	0	0
Похищение	X	0	0
Маг	0	X	0
Эльф	0	X	0
Гном	0	X	0
Орки	0	X	0
Оператор	0	0	X
Цикл	0	0	X
Процедура	0	0	X

Таблица 3

Матрица B	Собака Баскервилей	Пляшущие человечки	Властелин Колец	Хоббит	Искусство Програм-мирования	Программа = Алгоритм + Структура данных
Dim1	X	X	0	0	0	0
Dim2	0	0	X	X	0	0
Dim3	0	0	0	0	X	X

Если усовершенствовать синтаксическую модель предложения с учетом других возможных синтаксических позиций слов, то размерность модели увеличится. Скажем, структура «подлежащее – сказуемое – дополнение – определение – обстоятельство» для записи сочетаемости потребует массива размерности 5. Для обобщения модели считаем, что имеем дело с N-мерным массивом.

Методика сборки N-мерного текстового корпуса

Сначала текстовый корпус проходит этап синтаксического анализа предложений текстов, который производится с помощью Стеэндфордского парсера Stanford Parser [7].

Далее, разбирая синтаксическое дерево, постсинтаксический анализатор выделяет главный глагол предложения – на него указывает root (ROOT-0, ***verb***); субъект-подлежащее – nsubj (verb, ***noun***); прямой объект-дополнение – dobj (verb, ***noun***); непрямой объект – iobj (verb, ***noun***); существительное в предложной группе – prep_during (verb, ***noun***), prep_on (verb, ***noun***), prep_in (verb, ***noun***) и т.д.; межглагольную связку xcomp (verb,***verb*1**). Таким образом, при анализе предложения, находя лексемы в соответствующих синтаксических позициях, система заполняет этими словами кортеж предложения (root-verb, nsubj, dobj, iobj, prep_, xcomp, count) , при этом в prep_ существительное записывается вместе с предлогом. Если в предложении отсутствует некоторая синтаксическая позиция, то она заполняется символом пустого слова ***0***. В шестимерный массив данных помещаются только кортежи с как минимум тремя ненулевыми полями. В count сохраняется число раз использования подобного лексического сочетания в данном корпусе. Шесть первых элементов кортежей формируют координаты пространства, седьмой – значения частоты сочетаний. Как результат формируется 6-мерный массив сочетаний слов в данных синтаксических позициях предложений текстов корпуса.

Разложение тензора

Полученный массив – тензор размерности 6 – должен быть разложен в виде шести матриц, каждая из которых будет представлять отображение множества лексем, стоящих в определенной синтаксической позиции, на множество *k* фактор-измерений латентного семантического пространства семантико-синтаксических отношений слов текстового корпуса.

Для разложения тензора используется метод неотрицательной тензорной факторизации. Он подобен параллельному факторному анализу с ограничением, что все данные должны быть неотрицательными. Параллельный факторный анализ – это мультилинейный аналог сингулярного разложения матриц, используемого в латентном семантическом анализе. Главная идея метода – минимизация суммы квадратов разниц между оригинальным тензором и факторизированной моделью тензора. Для *N*-арного тензора $T \in R^{D_1 \times D_2 \times .. \times D_N}$ определяется целевая функция (1), где *k* – размерность факторизированой модели, а∘ – внешнее произведение (outer product).

$$\min_{x_i \in R^{D_1}, y_i \in R^{D_2}, \ldots, z_i \in R^{D_N}} \| T - \sum_{i=1}^{k} x_i \circ y_i \circ \ldots \circ z_i \|_F^2 . \qquad (1)$$

Для неотрицательной факторизации добавляются ограничения по неотрицательности значений элементов (2):

$$\min_{x_i \in R_{\geq 0}^{D_1}, y_i \in R_{\geq 0}^{D_2}, \ldots, z_i \in R_{\geq 0}^{D_N}} \| T - \sum_{i=1}^{k} x_i \circ y_i \circ \ldots \circ z_i \|_F^2 . \qquad (2)$$

Результат работы алгоритма – представление тензора в виде *N* матриц, которые описывают отображение каждой из размерностей тензора на *k* фактор-измерений латентного семантического пространства. Обычно NTF модель подгоняется методом наименьших квадратов. На каждой итерации *N-1* размерность фиксируется, а *N*-я размерность подгоняется методом наименьших квадратов. Процесс продолжается до момента сходимости. Число фактор-измерений латентного семантического пространства было взято *k* = 150. Исходя из опыта предыдущих исследований, именно это значение обеспечивает лучшие результаты факторизации [2]. Для решения данной задачи была написана программная реализация алгоритма параллельной факторизации PARAFAC [8] 6-мерного тензора, где значительного ускорения процесса решения задачи удалось достичь благодаря распараллеливанию вычислений на графической карте по технологии, аналогичной описанной в [9].

В результате факторизации собранного шестимерного тензора получены шесть матриц, состоящих из *k*-мерных векторов-столбиков и представляющих отображение множества слов в шести разных синтаксических позициях на *k*-мерное пространство сочетаний слов в предложениях корпуса. Для того чтобы вычислить частоту сочетаний слов **a b c d in_e f** в синтаксической последовательности NSUBJ, VERB, DOBJ, IOBJ, PREP_, XCOMP нужно вычислить сумму:

$$\sum_{i=1}^{k} \mathrm{NSUBJ[a]} * \mathrm{VERB[b]} * \mathrm{DOBJ[c]} * \mathrm{IOBJ[d]} * \mathrm{PREP_[in_e]} * \mathrm{XCOMP[f]} .$$

Таким образом, получено средство удобного представления шестимерного многомерного массива и быстрого эффективного вычисления значений данного массива. Можно легко вычислить частоту сочетаний типа «Электрик прикрутил лампочку», «Повар зажарил утку», «Поезд выехал в Симферополь» в текстах корпуса. Для этого нужно найти вектора-столбики, соответствующие данным словам из матриц, соот-

ветствующих их синтаксическим позициям, и вычислить сумму из произведений их координат. Используя данный инструментарий, можно различать правильные последовательности слов от неправильных, например, что «охотник подстрелил зайца» – это корректное высказывание, а «заяц подстрелил охотника» – нет. Данные матрицы разложенного многомерного тензора представляют собой латентное семантическое пространство семантико-синтаксических связей слов естественного языка. Семантико-синтаксическое поведение каждого слова, те связи, которые оно образовывает с другими словами, неявно описывается в его k-мерных векторах-столбцах в разных матрицах разложенного тензора, соответствующих его возможным синтаксическим позициям. Назовем эти k-мерные вектора-столбцы **семантико-синтаксическими валентностями** слов.

Заметим, что k-мерные векторы-столбцы из матриц разложенного тензора являются описанием частотного распределения лексем в последовательностях слов предложений. Основная сложность состоит в том, что при построении массива семантико-синтаксической сочетаемости лексем основными объектами изучения и анализа являются лексемы – слова, которые по природе неоднозначны. И векторное представление семантико-синтаксических валентностей любого слова W, коим является соответствующий ему вектор-столбец из матриц разложенного тензора, – это есть, по сути, сумма составляющих слагаемых векторов отдельных разных семантических значений этого слова W – концептов Sw_1, Sw_2,…, Sw_t в некоторой онтологии. Таким образом, стоит задача по вектору валентности $(v_1,v_2,\ldots,v_k)$ некоторого слова W получить составляющие слагаемые векторы валентностей $(v_{11},v_{12},\ldots,v_{1k})$, $(v_{21},v_{22},\ldots,v_{2k})$,.., $(v_{t1},v_{t2},\ldots,v_{tk})$ для каждого из его t значений. Вектор валентностей фиксированного значения – концепта некоторой онтологии – является неявным описанием его семантических отношений с другими концептами данной онтологической базы знаний. В работе описан метод определения семантических отношений между концептами – синсетами Wordnet, посредством анализа разложенных тензоров, сформированных при обработке корпусов статей English Wikipedia, с расщеплением векторов семантической валентности слов на составляющие вектора семантической валентности их значений, и с конкретной привязкой расщепленных векторов к соответствующим концептуальным узлам онтологии WordNet.

Алгоритм расщепления векторов семантико-синтаксической валентности слов на составляющие слагаемые векторы валентностей их разных значений

После факторизации шестимерного собранного тензора корпуса статей English Wikipedia были получены шесть матриц ROOT VERB, NSUBJ, DOBJ, IOBJ, PREP , XCOMP, которые состоят из векторов размерности k. Каждый вектор-столбец этих матриц соответствует некоторому слову или словосочетанию. Данные вектора описывают семантико-синтаксическое поведение слов, а именно, в каких синтаксических позициях какие связи и с кем некоторое слово образует. По аналогии с химической терминологией, назовем данные вектора **векторами семантико-синтаксических валентностей (VSV)** слов. Слова по своей природе являются неоднозначными, то есть им, как правило, соответствует несколько значений. Таким образом, вектор слова является суммой векторов всех значений данного слова. Одному слову может соответствовать несколько векторов из разных матриц, соответствующих разным синтаксическим позициям, задача расщепления каждого из этих векторов решается отдельно. Разработанный алгоритм расщепления **VSV** слова на множество **VSVs** всех его значений – синсетов Wordnet – имеет следующий вид:

Дан вектор семантической валентности *V*, размерности *k*, который соответствует некоторому слову *w* в NSUBJ (или же в любой другой из 6 матриц – метод работает аналогично). Существительному *w* соответствует *t* значений – синсетов в WordNet. Требуется разделить *V* на составляющие слагаемые $V_1, V_2, \ldots V_t$, соответствующие данным *t* синсетам.

```
 for i=1 to k do
 begin
        if V[i]<>0 then
 for j=1 to t do
begin
```

{Необходимо определить, какому из t синсетов принадлежит i-тое значение V[i]. Оно может принадлежать либо одному из синсетов, либо нескольким – тогда нужно разложить V[i] на сумму $V_1[i]+V_2[i]+\ldots+V_t[i]$}

Для j-того концепта найти ближайшего соседа в сети Wordnet, содержащего в составе своего синсета такое слово word, которому в NSUBJ соответствует вектор с i-тым элементом, большим, чем некоторый пороговый уровень **Th**. Если в синсете несколько таких слов, то выбираем слово с наибольшим значением i-того элемента его вектора из NSUBJ. При поиске ближайшего соседа учитываются только связи-отношения типа гипернимия – гипонимия. При нахождении ближайшего соседа, удовлетворяющего данным условиям, запоминаем i-тый элемент вектора его слова, поделенный на расстояние от концепта j до него:

$$X[j]=\frac{\text{NSUBJ[word ,i]}}{\text{distance(concept(j) , closest neighbor)}};$$

end;

$V_1[i], V_2[i], \ldots V_t[i]$ определяются из системы уравнений:

1. $\mathbf{V_1[i]} = X[1] *Y$; $\mathbf{V_2[i]} = X[2]*Y$; … ;$\mathbf{V_t[i]} = X[t] * Y$;

2. $\sum_{j=1}^{t} X[j]*Y = V[i]$;

Определить значение $Y=\frac{V[i]}{\sum_{j=1}^{t} X[j]}$;

For j=1 to t do if X[j] *Y< ***R*** then $\mathbf{V_j[i]}$ =0 else $\mathbf{V_j[i]}$ = X[j]*Y;

{***R*** – пороговый уровень, подобранный экспериментально}

end;

Эксперименты с векторами семантической валентности концептов-синсетов WordNet, полученными в результате работы описанного алгоритма, показали высокую точность расщепления векторов семантической валентности слов на составные слагаемые вектора семантической валентности их значений – концептов, с привязкой их к конкретным синсетам WordNet. Оценка точности в среднем достигает 91 – 92 %. Для проведения экспериментов была разработана программа, генерирующая по набору полученных **векторов семантических валентностей** синсетов WordNet множество

всех возможных последовательностей слов – предложений с их участием, которые согласуются со значениями в этих *k*-мерных **векторах** синсетов. Далее эксперты с помощью той же программы провели анализ корректности сформированных словосочетаний и вычисление оценки точности определения вектора семантической валентности для каждого синсета WordNet из тестовой выборки.

Выводы

Данная работа рассматривает модель *N*-мерного лингвистического пространства семантико-синтаксических отношений между словами естественного языка, которое формируется в результате частотного анализа синтаксических структур предложений больших текстовых корпусов. Данные представляются в виде *N*-мерных массивов данных, которые потом обрабатываются методами неотрицательной факторизации *N*-мерных тензоров. Разложение собранных тензоров в виде N матриц сокращенной размерности *k*, помимо компактной и удобной структуры представления данных о сочетаемости последовательностей лексем в некоторых синтаксических позициях предложений естественного языка, дает эффективный метод вычисления оценки вероятности существования семантико-синтаксических связей между словами разных грамматических категорий. При этом можно рассматривать *k*-мерные векторы из матриц разложенного тензора как векторы семантико-синтаксических валентностей слов. Так как слова по своей природе являются неоднозначными и одному слову, как правило, соответствует несколько значений, в работе предложено рассматривать *k*-мерные векторы семантико-синтаксических валентностей слов как суммы составных слагаемых векторов разных значений этих слов. В статье представлен разработанный метод расщепления векторов семантико-синтаксических валентностей слов на составные слагаемые векторы их разных значений с привязкой этих расщепленных векторов к синсетам WordNet в качестве их собственных значений векторов семантической валентности. Реализованный алгоритм был протестирован проведением ряда экспериментов с матрицами разложенного тензора корпуса текстов статей Wikipedia. Полученные при тестировании оценки точности работы предложенного алгоритма демонстрируют его высокую эффективность и говорят о реальных перспективах его использования на практике в автоматизации методов наполнения контентом онтологических баз знаний для автоматического определения семантических отношений между концептами – узлами онтологической сети – в процессе обработки больших текстовых корпусов.

Литература

1. Tim Van de Cruys. A Non-negative Tensor Factorization Model for Selectional Preference Induction / Tim Van de Cruys // Journal of Natural Language Engineering. – 2010. – № 16 (4). – P. 417-437.
2. Tim Van de Cruys. Multi-way Tensor Factorization for Unsupervised Lexical Acquisition / Tim Van de Cruys, Laura Rimell, Thierry Poibeau and Anna Korhonen // Proceedings of COLING 2012. – Mumbai, India. – P. 2703-2720.
3. Shay B. Cohen, Michael Collins. Tensor Decomposition for Fast Parsing with Latent-Variable PCFGs. NIPS 2012: 2528-2536
4. Wei Peng. On the equivalence between nonnegative tensor factorization and tensorial probabilistic latent semantic analysis / Wei Peng; Tao Li // Applied Intelligence, Springer Journals. – 2011. – October. – Volume 35, Issue 2. – P. 285-295.
5. Miller G.A. WordNet: An online lexical database / G.A. Miller, R. Beckwith, C.D. Fellbaum [and other] // Int. J. Lexicograph. – 1990. – № 3, 4. – P. 235-244.

6. Scott Deerwester, Susan T. Dumais, George W. Furnas. Indexing by Latent Semantic Analysis (PDF) / Scott Deerwester, Susan T. Dumais, George W. Furnas [and other] // Journal of the American Society for Information Science: – 1990: – 41 (6): – P. 391-407.
7. [Электронный ресурс]. – Режим доступа : http://nlp.stanford.edu/software/lex-parser.shtml
8. Harshman R. Foundations of the PARAFAC procedure: models and conditions for an explanatory multimodal factor analysis / R. Harshman // UCLA Working Papers in Phonetics. – 1970. – № 16. – P. 1-84.
9. Jukka Antikainen. Nonnegative Tensor Factorization Accelerated Using GPGPU / Jukka Antikainen, Jirí Havel, Radovan Josth [and other] // IEEE Trans. Parallel Distrib. Syst. 2011, 22(7): pp. 1135-1141.

RESUME

O.O. Marchenko

Algorithm for automatic detection of semantic relations between concepts of ontologies

This paper describes a method for automatic determining semantic relations between concept nodes of an ontological knowledge base by analyzing the matrices of words semantic-syntactic valences. These matrices are obtained by non-negative factorization of tensors of words syntactic combinability. The tensors have been generated in the course of syntactic structures frequency analysis for the large text corpus of English Wikipedia articles.

The precision assessment of the proposed algorithm, obtained during the performed testing, demonstrates its high efficiency and provides the real prospects for its practical use in automation of methods for filling the content of ontological knowledge bases to automatically determine the semantic relations between the ontological network nodes in the processing large text corpora.

Статья поступила в редакцию 03.04.2013.